DOCUMENT RESUME

ED 161 320 HE 010 527

AUTHOR Sturges, Jack
TITLE How to Make the Most Out of Course Evaluation Forms.
PUB DATE 78
NOTE 25p.; Paper presented at the Educational Innovations Exchange, Council on Social Work Education Annual Program Meeting (New Orleans, 1978)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS College Faculty; *Course Evaluation; Educational Assessment; Effective Teaching; *Evaluation Methods; *Faculty Evaluation; Higher Education; Literature Reviews; Measurement Techniques; Questionnaires; *Rating Scales; Reliability; Statistical Analysis; *Student Evaluation of Teacher Performance; Teaching Quality; *Validity

ABSTRACT

The questionnaires or course evaluation forms (CEF) are routinely administered to university students to obtain information about the quality of courses and instructor effectiveness. There is an extensive variety of formats but many use a combination of Likert-type items and open-ended questions. Percentages, but more often means and standard deviations, are used to report student response to CEFs. A review of the literature reveals that CEFs may not elicit accurate information from students concerning course quality because there are a number of extraneous variables that influence student responses. There is some question about the accuracy of information obtained by treating nominal scale data with statistical techniques designed for interval or ratio scale data. Some schools of social work are administering CEFs of unknown reliability and validity. A procedure is presented for developing a functional CEF, and suggestions are offered for determining whether differences in student ratings of course quality are significant. It is concluded that defensible and usable data will result only from CEFs that possess demonstrable reliability and validity, and the CEFs of value must yield data that faculty members will understand, accept, and apply to course and curriculum development. (SW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY Jack Stringer

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

ED161320

HOW TO MAKE THE MOST OUT OF COURSE EVALUATION FORMS

"Students are a prime source of information about the value of courses." "Students don't have the sophistication to judge the worthwhileness of a course." "Course rating forms are really coarse rating forms." "Course rating forms have been proven to be reliable and valid." "Have you ever seen any real changes in teaching because of student ratings or comments?" "If you ask students to do very much work your ratings will be low." So go the comments heard from social work educators more often now that course evaluation forms are being used almost everywhere.

Introduction

Social work educators are now regularly asking students for information about the value of the courses they have completed. A common way of obtaining this information is to ask students to record their opinion of the quality of courses on questionnaires, i.e., course evaluation forms (CEF's). The purpose for conducting such inquiries is to obtain information about the quality and effectiveness of social work courses, instructors, and the curriculum. Questions are asked about courses because accrediting agencies, university administrators, social work practitioners, and social work educators need to know about course quality. Afterall, it is essential to evaluate the quality of courses in order to develop a curriculum that will prepare individuals for professional practice.

The first thing noticed in looking at CEF's is that they often contain questions about both the quality of the course and about the effectiveness of the instructor. However, it is difficult to determine whether questions on CEF's concerning the quality of a course are distinct or really different from those about effectiveness of the instructor. By the variety of CEF's presently in use, it appears that social work educators are constantly involved in modifying old CEF's or developing new ones. This rather constant (at least annual) attention to

HE 010 527

developing new CEF's seems to result from some dissatisfaction (or distrust) of those previously used.

Whatever the form of CEF's in present use and whatever the attitude toward those CEF's, it is apparent that they are now a regular part of the evaluation of social work education programs. Unfortuantely, there seems to be some confusion, differences of opinion, and even disagreement, about what the format of CEF's should be, what questions students should be asked, and how student judgment about courses might be most efficiently obtained and interpreted.

Related Research and Literature

A review of the literature concerning evaluation of teaching effectiveness, course evaluation, and CEF's shows that so much has been written, a thorough analysis would be a most formidable task. Even a brief review, however, can provide a substantial amount of information about a surprisingly large assortment of techniques, procedures, and questionnaires used in the evaluation of social work courses and programs. In spite of the array of procedures, it is clear that the overall purpose of course or program evaluation is to improve the quality of instruction.

Evaluation techniques commonly used in schools of social work include evaluations by colleagues, administrators, students, social work practitioners, and self-evaluation. Self-evaluation is a process by which instructors rather informally assess the quality of their teaching and the content of courses they have taught. At some institutions, the results of self-evaluation become a part of a formal faculty evaluation procedure but more often the results are only casually shared with others and informally used by the instructor in modifying teaching behavior and course content. Colleague evaluation is a process by which faculty members provide their judgment and information about the quality of courses and the teaching actitivities of their colleagues. Data used in colleague evaluations range from ratings obtained through faculty visiting the classrooms of their colleagues to rather subjective information obtained through general impressions, student comments, and the like. Practitioner evaluations of course quality are usually obtained from social work practitioners who supervise students enrolled in educational practicums and from those who employ graduates of a program. Evaluations by administrators are generally based on information obtained from faculty self-

reports, students, alumni, faculty, and personal observation of courses and instructors. Student evaluations often consist of data concerning the quality of courses and instructor performance obtained by CEF's administered during the last few days of the academic term. While each of these methods of evaluation has limitations, each can be useful in evaluating the quality of courses offered in a social work education program.

The technique used most often for course evaluation is that of giving students CEF's and asking them to both rate and to make statements about the quality of courses and instructors. The arguments for the use of student judgment is that students are capable of judging the quality of their education, they are the only ones exposed consistently to the process, and as "consumers," are the best source of information. It is also argued that the use of CEF's provides a standardized way for all students to evaluate courses, that all courses can be evaluated at relatively little cost, and that CEF's are easy to administer and yield normative data.

The literature concerning course evaluation provides some information suggesting that data obtained from students about the quality of courses and teachers are as accurate and dependable as data obtained from other sources. For example, Costin, Greenough, and Menges (1971) and Miller (1972), report that the research they reviewed led them to believe that 1) student ratings can provide reliable and valuable information, 2) student ratings can have acceptable stability and internal consistency, 3) student ratings of instruction are related to the quality of instruction rather than to entertainment value or personality characteristics of instructors, and 4) data obtained from students' responses to CEF's may aid instructors in improving course content and teacher effectiveness.

Carlson and Park (1976) report that if CEF's have been well-constructed "there is increasing evidence that students are capable of making fair and informed judments of a teacher's performance." Baliski (1976, p. 26), Cook and Neville (1971), Doyle (1975), Miller (1974), Pyper (1971), and Stake (1976) all speak to what a well-constructed CEF should be like and report favorably on CEF rationale and usability. Doyle (1975), Grinnell and Kyte (1976), Millard

Millard (1976), Miller (1972), and Miller (1974) give examples of components of courses that can be assessed by CEF's. Miller, (1974, p. 35) warns, however, that student evaluations provide good data only when they are developed by faculty deciding just what is to be evaluated.

While Thomas (1976) notes that there is the question of what criteria of course value should be used in assessing different aspects of course content and teaching effectiveness, Perry (1969) reports an agreement among students about what they see as important teacher behavior and course quality. Further, it seems that student ratings of instructor effectiveness may be better than self-ratings by instructors because instructor ratings of their own effectiveness tend to be inflated (Blackburn and Clark, 1975; Centra, 1972; Clark and Blackburn, 1973; Sorey, 1967).

Additional evidence concerning the validity of student responses to CEF's is provided by Aleamoni and Yimer (1973), Cohen and Berger (1970), Faia (1976), and Swanson and Sisson (1971). These studies generally report a relationship between student ratings to rating of colleagues, expert judges, and to some degree, to the amount learned by students. Other research has also addressed the validity of CEF's by assessing variables suspected of influencing students judgment of course quality. Variables reported to be non-influencing or of little influence on the validity of student ratings are student GPA (Aleamoni, 1976), size of class (Aleamoni and Graham, 1974; Costing et al., 1971), anonymity of student respondent (Abrami, Leventhal, Perry, and Breen, 1976), sex of student (Aleamoni, 1972; Costin et al. 1971), or time of the day the course was offered (Yongkittikul, Gilmore, and Brandenburg, 1974). Other data concerning a particular CEF, The Course and Instructor Evaluation Questionnaire (CIEF), lead Aleamoni (1977) to state that CEF's can be reliable and that they can yield data that do indeed identify courses that are either very poor or very good. He contends that it is only the quality of course content and instructor performance that determine student ratings of the value of a course. In short, there are a substantial number of statements by authorities and reports of research which provide support for the use of CEF's in evaluating quality of instruction and performance of instructors.

In spite of the number of statements in support of CEF's, there are statements and arguments about the shortcomings of CEF's. For example, some faculty say that teaching is an art and cannot be measured by CEF's. Some feel that the value of a course cannot be judged by naive judges such as students. Other faculty state that the true value of a course cannot be determined by students until they have had a chance to apply their knowledge in later "real life" situations. Others say that CEF's are of little value since style of CEF administration affects student responses. Further, items on CEF's are not applicable to courses in field instruction or labs, students' perceptions are influenced by a variety of unidentifiable factors, data from CEF's are neither reliable nor valid, and all of the problems normally attending questionnaires are associated with CEF's. Some faculty show little interest in student opinion about course quality because they feel that the amount of research and the number of publications of instructors are the real measures of the value of courses.

There is research that suggests that CEF's produce data that are inaccurate or that are of questionable value. Isaacson, McKeachie, and Milhollan (1963) were early reporters of how differences in ratings of teacher effectiveness might be due to teacher personality, credibility of lecturer, size of class, level of students, and the like, rather than to quality of instruction. Aleamoni and Graham (1974), and Costin et al. (1971), while generally supporting the accuracy of CEF data, report that some differences in course evaluation may result from instructor rank, class size, the course being required or elective, etc. Brown (1976), Crawford and Bradshaw (1968), Sheehan (1975), and Wilson, Gaff, Dienst, Wood, and Bavry (1975), report data indicating that student GPA, instructor friendliness outside the classroom, instructor personality differences, etc., affect students' responses to CEF questions about course quality and instructor peformance.

Others who report variables which seem to to affect student responses to CEF's are Elliot (1949), Medley (1973), and McKeachie, et al. (1971). They state that their data suggest that attributes of effective teaching are not considered to be the same by male and female students, that teachers effective with high ability students may not be as effective with low ability students, and that students may learn as much

from teachers rated ineffective as they do from teachers rated effective. Dennis (1976) points to evidence that many of the traits that students see as characteristic of good teachers have nothing to do with scholarship or good teaching. In fact, the manner of presenting the material of a course may be seen as a greater value to students than the actual content. Starkman (1977) concurs and explains that while course evaluations are often made by obtaining the collective sentiments about courses, not everyone sees course effectiveness in the same way.

A variety of other factors seem to influence the manner in which students rate the value of a course. Faia (1976), Sheehan (1975), and Sherman (1976) relate how instructors can manipulate student ratings on CEF's. Abrami, Leventhal, Perry, and Breen (1976), and Gupta (1976), provide information about how the sponsorship of CEF's can also affect student evaluations. Sullivan and Skanes (1974) state in their research they found only a small-to-modest association between student achievement and student evaluation of instructors. Doyle (1976, pp. 38-58) says that student leniency (kindness) affects the ratings they give to instructors, and that a halo effect can influence ratings. He also suggests that students might rate highest those instructors from whom they have learned the least. Kirchner (1969) reported that in his research, student ratings were higher when instructors who were being rated were in the room than when they were absent. Collivar (1972) and Stone, Robinowitz, and Spool (1977) report data showing that anonymity of student response can affect the students' ratings of a course or an instructor. Stone et al. (1977) also suggest that positiveness or negativeness of the wording of CEF items will affect the ratings given by students.

DeYoung (1977) says that the overall classroom climate affects course evaluations. Centra (1973) suggests that it is difficult to identify good teacher and/or good courses because in most institutions all teachers and their courses tend to receive reasonably good ratings from students. Further, Abrami, Leventhal, Perry and Breen, (1976) present data that suggest that students who select courses on the basis of the teacher's reputation will rate those courses as being better than courses rated by students who did not choose courses on that basis. Gordon and Hudson (1977) however, indicate that some groups of students simply cannot agree about whether the course or the

instructor was excellent or extremely poor.

A major criticism aimed at CEF's is that the data they produce sometimes result from responses to CEF questions that, according to Harari and Zedeck (1972), are ambiguous, arbitrarily chosen, and inapplicable in some courses. Dressel (1976, p. 349) also states that the interpretation of CEF results become difficult when instructor or course norms were derived from responses to questions which were selected to assess a variety of courses. He mentions that the instructor rated low on such norms is naturally inclined to criticize the instrument or the procedure rather than improve. High scores on the other hand would tell the instructor that no improvement was needed. It would seem then that student ratings of instructors and courses, especially if accompanied by inappropriate norms, could actually cause faculty to resist making instructional improvements.

The literature reports a variety of CEF formats. Some ask for students to write essay kinds of evaluations, some ask for response to a type of graphic rating scale or a type of Semantic Differential scale, some ask for agreement or disagreement with statements about the course on a modified Likert-type scale, and most are some combination of these (See Anastasi, p. 551, pp. 602-612, 1976; Likert, 1932; Osgood, 1957 for scale descriptions.) Many CEF's allow for ratings of both the course and the teacher (usually on a Likert-type scale), written comments about whatever the respondent feels is important to point out, or a reporting of "critical incidents" exemplifying the relative effectiveness of the course or the teacher. Aleamoni and Spencer (1973), Doyle (1975), McKeachie (1969), Miller (1972), and Miller (1974), for example, present a variety of CEF's, most of which employ a Likert-type scale.

Much of the research concerning CEF's reports data derived from responses to Likert-type scales and the data are often reported in means, standard deviations, Pearson r's, student's t's, factor analytic values, and other statistics that can be computed from interval or ratio scale data. There may be cause for concern here since some measurement experts say the Likert scales yield frequency, or at best ordinal data, and the treatment of such data as if they were ratio interval level could lead to faulty interpretations. Since this concern

seems to be rather infrequently addressed in the literature, it might be that faculty members responsible for developing CEF's have not been consulting regularly with their statistician colleagues.

Another problem of data interpretation is that reliability coefficients often reported for CEF's are computed from "split-half" techniques for determining correlation. These techniques yield an index of the internal consistency, i.e., homogeneity, or "going togetherness" of the items or ability of Likert-type scales to consistently obtain the same or similar information from respondents. What these "split half" techniques do provide is an estimate or index of whether each, or some combination, of the item on a Likert-type student rating form seem to be assessing the same domain of attitude or judgment about the course. They do <u>not</u> provide real evidence about the degree to which the instrument can elicit stable or consistent responses from students about their judgments of a course or the instructor.

This matter of obtaining solid evidence about the reliability of CEF's is crucial since the ability of an instrument to obtain accurate, i.e., valid information from students is directly related to the instrument's ability to obtain consistant, i.e., reliable, information. Cronback (1970, p. 171) points out validity [accuracy] coefficient of an instrument cannot exceed the square root of the reliability [stability] coefficient of that instrument. So, an instrument of little ability to obtain consistent information from students has little ability to obtain accurate information from students. It follows that CEF's of unknown reliability yield data of unknown validity. The literature concerning validity of CEF's is very limited which further supports the notion of unsupported validity of many CEF's in current use. Meeth (1976) contends that it is strange that a profession as important as teaching can continue . . . . on such shallow basis of evidence of effectiveness.

Educators might note that using CEF's possessing unknown validity could become a serious matter if the data are used in making decisions about instructor hiring or promotion. The reason being that most CEF's probably meet the definition of "test" as stated by the Equal Employment Opportunity Commission (EEOC) and that the use of tests which could adversely affect hiring, promotion, etc., could run counter to the EEOC "Guidelines on Employee Selection Procedures" (1970). Even if it

is curriculum changes and not personnel decisions that are made on the basis of data from CEF's, the changing of teacher behavior or curriculum content on the basis of faulty data would certainly be no trivial matter.

In summary, the review of the literature indicates that CEF's are commonly used to obtain information from students about the quality of courses and teacher performance. There is, however, disagreement about whether student opinion is dependable and whether CEF's yield reliable and valid data. Argument for and against the use of CEF's abound. A variety of CEF formats exist and many yield frequency data which is treated as interval or ratio scale data. The literature indicates that while some institutions have evidence of reliability and validity of their own forms, CEF's of questionable reliability and validity are used routinely. Apparently, there are many instances where appropriate, or at least optimal, use of CEF's does not exist.

Making the Most Out of CEF's

In order to use CEF's to their greatest advantage, a series of actions can be taken to increase the likelihood that CEF's will yield data that will be useful in improvement of course content or teacher performance. These steps call for the faculty to:

1) Define the goals and objectives of the education program.
2) Determine what curriculum content and learning experiences are necessary to meet program objectives.
3) Determine which courses need to be offered and state the specific objectives, goals and content of each and show how each course relates to other courses and to overall program objectives.
4) Define the teaching behaviors and the learning experiences for each course that are necessary to ensure that students meet the objectives of the course.
5) State in detail the student behaviors that would constitute evidence of mastery of course objectives and develop techniques for obtaining this evidence.
6) Develop a purpose, a rationale, and techniques for assessing the quality of course content and teacher effectiveness.
7) Develop CEF's that contain items directly related to course objectives, student learning, and teacher effectiveness and provide evidence of reliability and validity.

8) Develop standard techniques for administering and scoring the CEF's and for reporting and interpreting the data.
9) Ensure that faculty have been involved in CEF development and that they will understand, accept and apply CEF results.
10) Establish procedures for including CEF data as part of the overall procedure for evaluating and planning curriculum.
11) Develop a standard procedure for continuous evaluation of the course evaluation program.

In explanation of the above, it seems quite logical that any evaluation of quality of instruction would be preceded by an explanation of the goals and objectives of the instructional program. Afterall, it is difficult to assess the quality of something that has been produced if the characteristics and attributes of the intended product have not been described before production began. Further, instructional goals need to be consistent not only with requirements for professional practice but with University objectives, characteristics of the student body, and with the requirements of professional accrediting agencies. Also, the objectives and the content of the curriculum need to be derived from the specific knowledge, values, and skills necessary for the professional practice of social work. A social work education program must also identify how the different components or areas of knowledge of the curriculum can be presented in the most effective way to achieve curriculum objectives. Each course activity, unit, etc., offered should be developed so that not only can it meet the curriculum objectives for which it was designed but so that it is related or integrated with the other courses, units, activities, etc.

In order to more easily assess the quality of a course, course objectives can be stated in behavioral terms, i.e., in terms of what students will be able to do (and under what circumstances) if they have mastered the content of the course. While stating course objectives in behavioral terms is not the only way of stating objectives, there needs to be a place in course construction for precise statements

about which practice skills or knowledge are to result from which content, which courses, and from which learning activities. The difficult task is to state which student behaviors will constitute evidence of mastery of content or objectives and to then develop techniques for observing and measuring those behaviors. The idea is to be able to compare precisely stated course objectives with specifically described student behaviors and arrive at a measure of course effectiveness.

One way of determining whether courses are effective in preparing students for practice is to simply ask their opinion about the quality of their education. Without supporting information, the difficult-to-defend assumption here is that student opinion of the quality of courses or instructors is related to amount of learning. Nevertheless, student opinion can be valuable and methods for obtaining reasonably accurate student judgment of course quality can be developed. While student opinion can be obtained through interviews, class sessions devoted to critiquing the course, etc., the method which seems to have the widest use is to ask students to respond to a locally developed CEF.

The first step in developing a functional CEF is to decide just what the CEF is to measure. It must be determined which domains of instructor behavior and course content should be measured by the CEF. Usually, the domains or "traits" of courses that need to be assessed are course content, course organization and materials, and teacher performance. Therefore, questions or statements such as "Was the content timely and relevant?", "Were the lectures presented in an organized fashion?", and "Was the instructor enthusiastic?" appear on CEF's. In order for a CEF to provide data of any real value, there needs to be strong agreement among faculty about what domains of content and behavior should be assessed and about what the format of the CEF should be.

Once faculty agree upon just what it is that the CEF is to assess, a variety of questions, statements, etc., which might be posed to students can be assembled. From this "pool" of possible CEF items, those which appear to be overlapping or duplicating others can be eliminated. The remaining items might be subjected to faculty and

student analysis designed to select those which are the most appropriate. Expert judges, psychometrists, and social work practitioners might also be of assistance in selecting items, areas to be assessed, and CEF format.

Once the domains to be assessed and the format is determined, the reliability of the CEF must be addressed. The question of CEF reliability is generally that of, "Does the CEF have the ability to consistently measure student opinion about the quality of a particular course?" One index of reliability can be obtained by having students respond to a CEF and then by having them again respond on a later date so that responses can be compared. The greater the agreement between the two sets of opinions, the greater the reliability, i.e., stability, of the instrument. Another way of looking at reliability is to see whether students respond consistently to all CEF items concerning certain areas of course quality or teacher behavior. The technique here is to compare student responses to items designed to measure opinion about a certain element of a course with their responses to other items designed to measure opinion about the same element. A strong agreement would indicate that the CEF possessed reliability, i.e., internal consistency.

A number of texts, e.g., Ghiselli (1964, p. 207), Guion (1965, p. 27), Kerlinger (1973, p. 442) explain techniques for determining reliability. While the work involved in developing a CEF of acceptable reliability can be substantial, the construction of a reliable CEF is essential if the data yielded by the CEF are to be given any weight in the evaluation of quality of course or instructor.

Most important in the construction of CEF's is the determination of whether they possess validity. The question of validity is generally that of "Does the CEF have the ability to accurately assess student judgment of the quality of the course?" Validity means the ability of the instrument to yield data that represent the real, true, or accurate judgment of students about course quality, instructor performance, etc. Obviously, a CEF which can provide accurate measurement of student judgment is necessary. However, care must be taken in interpreting even the most accurate student opinion of course or instructor quality as being evidence of the amount of course content actually

mastered by the students. Indexes of validity should represent not only the ability of the CEF to accurately obtain student opinion about a course but also the ability of the CEF to reflect the association between student judgment of course quality and the amount of course content mastered.

There are several ways of determining validity. One common way is to ask social work educators to examine the CEF and determine whether the items on the CEF represent the domains of course content and instructor behavior that need to be assessed. Another procedure is to compare student opinion of the effectiveness of the course with the degree to which they exhibit master of course objectives. Methods used to determine validity are explained by Anastasi (1976, p. 134), Cronbach and Meehl (1955), Ghiselli (1964, p. 335) Kerlinger (1973, p. 456), and other authors of texts in research and measurement. Aleamoni and Spencer (1973), Costin, et al. (1971), Doyle (1975, p. 47), and Miller (1974, p. 71) provide descriptions of how the validity of selected CEF's was developed.

The method of determining reliability and validity will be related to the format or type of construction of the CEF. The CEF may, for example, contain items which are considered discrete or distinct from any of the other items. The relative frequency of student course ratings or judgments of course quality for each item is considered to provide a measure of a particular or distinct course or instructor trait. These CEF items or questions produce nominal or ordinal scale data. Reliability may be determined by assessing the degree to which each item obtains consistent information from students. Validity can be determined through an item-by-item comparison of judgment of course quality with whatever criterion measures have been selected. The relative quality of a course or an instructor can be judged by comparing student responses to each item with their responses to each item with their responses to the same items in other courses. Agreement among student ratings, median ratings, and other nonparametric statistics can be computed to provide additional information. This method of determining reliability and validity and for evaluating the quality of courses on an item-by-item basis can be quite useful. The problem is that considering each item as being

a discrete measure and yielding only nominal scale data can result in making so many comparisons among traits and across courses that course evaluation can become lost in the mass of data and comparisons. Further, considering responses to CEF's as being nominal scale data precludes the application of more powerful parametric statistical analysis of the data.

Quite often, CEF's consisting of a number of items are subjected to analysis to determine whether the instrument yields a single overall measure of course quality. This procedure is a type of item analysis and calls for comparing the mean rating of each CEF item to the mean of all of the items. Items on which student ratings of the course do not agree with their overall or average ratings are considered to be measuring something different from the other items and are discarded. As Ghiselli (1964, p. 184) points out, ". . . eliminating those components with the lowest correlation with the total composite score results in increasing the homogeneity of the composite scores. . ." Validity may be determined by comparing the overall rating of course quality of the different courses to whatever criteria of effectiveness have been selected. Comparisons of quality of courses can be made by comparing mean or summed student ratings of each course. Since there is a high degree of association among items, it becomes difficult to identify specific teacher behaviors or areas of course content and organization that might be relatively better than others.

Data from CEF's can be subjected to a statistical analysis to determine whether the items might be obtaining information from students about distinct or separate components of a course. A common statistical technique for identifying which components are factors being assessed by the different items is generally referred to as factors analysis. This method has been borrowed from the field of psychological measurement where it was developed for the purpose of determining the number and nature of underlying factors which exist among a larger number of measurement of individuals. Anastasi (1976, p. 362) and Kerlinger (1973, p. 659), for example, discuss the application of factor analysis. The value of factor analysis is that it provides a way for determining whether there are one or more

clusters or groups of items on a CEF which seem to be measuring distinct areas or domains of course quality. Factor analysis provides for the discarding of irrelevant and extraneous CEF items or questions. Once the factors have been identified, each course can be provided with a total "score" and a "subtest score" for each of the factors identified. This allows for relatively easy comparison of total and factor scores of different courses. Also, individual instructors can examine student judgment of different factors of their own course and identify relative strengths of the different components. Reliability and validity of CEF factors is assessed by determining whether students respond consistently to the items comprising the different factors and by comparing factor scores to whatever criteria of course effectiveness have been selected.

The overall interpretation of means and standard deviations of student ratings of factors of a particular course is not difficult. However, the procedure for determining how individual instructors can determine whether there is any significant difference in the mean evaluation scores for their courses is rarely addressed. Usually, the data provided to instructors consist of mean, standard deviation, and number of respondents for each CEF factors. However, what is to be concluded about real differences in factor scores on the basis of, for example, mean scores of 3.2, 3.4, and 3.8 on factors of course organization, course content, and teacher effectiveness is not easy to state. An acceptable, but not widely explained, technique for comparing mean ratings of the factors is to estimate the amount of error of measurement in each of the factor means due to less-than-perfect reliability of the CEF and to then determine the fiducial limits of the means within which the "true" means would be if there were no error in the ratings. The next step is to determine which differences between means can be considered to be significant. The computation requires that the mean and standard deviation of student ratings of each factor for the course be provided and that the reliability (stability) coefficents of each of the factors be known.

The process to follow to determine whether the difference between any two factor means is large enough to consider the means as being significantly different is as follows:

Step 1: Compute the Standard Error of Measurement (SE_{meas}) for each mean factor score. The SE_{meas} provides an estimate of amount of error in the mean scores due to less-than-perfect reliability of CEF.

$$SE_{meas} \quad s\sqrt{1-r_{xx}}$$

where s = the standard deviation of the factor

and

where r_{xx} = The reliability coefficient (stability) of the factor

Step 2: Compute the Standard Error of Difference Between Factor Means (SE_{diff}). The SE_{diff} provides an estimate of amount of error in the difference between the factor means.

$$SE_{diff} = \sqrt{SE_{meas_1}^2 + SE_{meas_2}^2}$$

Step 3: Multiply the SE_{diff} times a selected alpha or confidence level value (usually 1.96). This will provide the difference between factor mean scores required to state whether there is probably a real difference between mean scores.

An example of how mean scores (derived from a 5-point Likert-type scale) may be compared is:

A)	FACTOR 1	FACTOR 2
	$r_{xx1} = .64$	$r_{xx1} = .84$
	$s_1 = .4$	$s_2 = .3$
	$\bar{X}_1 = 3.8$	$\bar{X}_2 = 3.4$
B)	$SE_{meas_1} = .4\sqrt{1-.64} = .24$	$SE_{meas_2} = .3\sqrt{1-.84} = .12$

C) $SE_{diff} = \sqrt{.24^2 + .12^2} = .26$

D) The required difference = alpha times SE_{diff} =

$$1.96(.26) = .51.$$

If the SE_{diff} times (1.96) is greater than the difference between two means, there is probably no real differences between the means. If SE_{diff} times (1.96) is less than the difference between the means, there is probably a real difference between the means. Since the difference between the means in the example is less than .51, the difference could have occurred because of error in the CEF data and probably not because of any real difference in student evaluation of the two factors of the course. (See measurement texts for detailed explanations of SE_{meas} and SE_{diff}.)

While this technique for comparing the mean scores of student ratings of the different components or factors of an individual course is useful, care must be taken in the interpretation if very small numbers of students were involved. Problems of interpretation also arise if the reliability coefficients were computed from data provided by students unlike those presently enrolled, or if there is considerable disagreement among students about course quality.

There is some difficulty in comparing student judgment of several courses because differences in student opinion of courses may result as much from differences between students enrolled in the different courses as from the differences in course quality. The assumption underlying comparisons of mean ratings of different courses is that students in each of the courses are from the same population with respect to education, age, expectation, or other factors which could affect their view of the course quality. Therefore, student judgment and/or factor scores concerning the quality of courses populated by groups of students which differ in any notable ways must be cautiously compared. If groups of students judging course quality can be considered to be from the same population with respect to extraneous variables that might systematically influence their evaluations, their ratings of courses can more readily be compared. Such comparisons might involve statistical tests related to means such as Student's *t*. (See texts in statistics, e.g. Winer, 1971, for assumptions underlying *t*-test for comparisons of means.)

Any interpretation of student opinion about the quality of individual courses or of individual instructors is difficult unless data showing student opinion about similar courses and similar instructors is available for comparison. Such normative data should be provided for

any sufficiently large, clearly definable group of students. Norm groups could be differentiated by level of course, difficulty of course, age and sex of students, previous work experience, instructor age or rank, and the like. Carefully developed normative data are necessary if the relative quality of a particular course is to be determined by comparing it to the quality of similar course.

The last step in making the most out of CEF's is to periodically evaluate the course evaluation program and to continue to provide data from course evaluation to the process of curriculum planning and course development.

Summary

The common practice now is to evaluate the quality of social work courses by asking students to respond to questionnaires or course evaluation forms (CEF's). These CEF's ask for student judgment about both the quality of the course and the performance of the instructor. The variety of CEF formats is extensive but many use a combination of Likert-type items and open-ended questions. Percentages, but even more often means and standard deviations, are used to report student responses to CEF's.

Some of the research shows that student responses to CEF's consist of accurate and valuable information about course quality and that this information can be applied toward improving the quality of instructional programs. Several Universities report that they have developed and successfully used rather sophisticated CEF's and that the instruments possess a considerable degree of reliability and validity. On the other hand, there are reports of research which show that CEF's cannot elicit accurate information from students concerning course quality because there are a number of extraneous variables which influence student responses. Further, there is some question about the accuracy of information obtained by treating nominal scale data with statistical techniques designed for interval or ratio scale data. Moreover, it appears that some Schools of Social Work are administering CEF's of unknown reliability and validity.

The procedure for developing CEF's that will yield usable results is to: (1) State curriculum and course objectives and develop a clear purpose and rationale for assessing course quality with CEF's; (2) Design CEF's that will yield data about the quality of teaching and about the quality of the separate components of courses; (3) Establish the reliability and validity of the CEF's and report normative data so that comparisons of courses and teachers can be made; (4) Instruct faculty in procedures such as Student's t and s_{diff} for determining differences between student judgment of different courses or components of individual courses; (5) Ensure that faculty have been involved in planning and developing CEF's and that they will understand, accept, and apply CEF results; (6) Continue to evaluate CEF's and to feed course evaluation data into the process of course and curriculum planning.

There is some doubt about whether CEF's can be developed which will provide precise measurement of the quality of courses or of instructor performance. There is also some doubt about whether the assumption that student assessment of course quality is related to the degree of mastery of course objectives is tenable. Nevertheless, CEF's can be developed which will provide more useful and dependable information than those presently being used in some Schools. If decisions are to be made on the basis of student responses to CEF's, it is imperative that the instruments possess demonstrable reliability and validity. If CEF's care to be of any real value, it is necessary that they yield data which faculty members will understand, accept, and apply.

REFERENCES

Abrami, P.C., Leventhal, L., Perry, R.P., & Breen L.J. Course evaluation: How? Journal of Educational Psychology, 1976, 68, 300-304.

Aleamoni, L.M. A review of recent reliability and validity studies on the Illinois course evaluation questionnaire (CEQ) (RM 127). Urbana, IL: University of Illinois, Measurement and Research Division, Office of Instructional Resources, 1972.

Aleamoni, L.M. Comments on the invalidity of student ratings for administrative personnel decisions. The Journal of Higher Education, 1976, XLVII, 607-611.

Aleamoni, L.M. The course and instructor evaluation questionnaire. Unpublished manuscript, Office of Institutional Research & Development, University of Arizona, 1977.

Aleamoni, L.M., & Graham, M.H. The relationship between CEQ ratings and instructor's rank, class size and course level. Journal of Educational Measurement, 1974, 11, 189-202.

Aleamoni, L.M., & Spencer, R.E. The Illinois course evaluation questionnaire: A description of its development and a report of some of its results. Educational and Psychological Measurement, 1973, 33 669-684.

Aleamoni, L.M., & Yimer, M. An investigation of the relationship between colleague rating, student rating, research productivity and academic rank in rating instructional effectiveness. Journal of Educational Psychology, 1973, 64, 274-277.

Anastasi, A. Psychological testing. (4th ed.). New York: McMillan, 1976.

Babski, C. Does students' expectations of teacher affect students evaluation of teachers? Miami, FL: Miami-Dade Community College, 1976. (ERIC Document Reproduction Service No. ED 118 206)

Blackburn, R.T., & Clark, M.S. Assessment of faculty performance. Sociology of Education, 1975, 48, 245-256.

Brown, D.L. Faculty ratings and student grades, a university-wide multiple regression analysis. Journal of Educational Psychology, 1976, 68, 573-578.

Carlson, R.V., & Park, R. Teacher evaluation: Relevant concepts and related procedures. Burlington, VT: University of Vermont, College of Education and Social Services, 1976. (ERIC Document Reproduction Service No. ED 129 739)

Centra, J.A. The student as godfather? The impact of student ratings on academia. Paper presented at the First Invitational Conference on Faculty Effectiveness as Measured by Students, Temple University, Philadelphia, 1973.

Clark, M.J., & Blackburn, R.T. Faculty performance under stress. In A. Sockloff (Ed.) Faculty Effectiveness As Evaluated By Students. Philadelphia: Temple University, 1973.

Cohen, S.H., & Berger, W.G. Dimensions of students' rating of college instructors underlying subsequent achievement on course evaluations. Proceedings of 78th Annual Convention of the American Psychological Association, 1970, 5, 605-606.

Collivar, J.A. A report on student evaluation of faculty teaching performance at Sangamon State University. A technical paper No. 1. Springfield, IL: Division of Academic Affairs, 1972.

Cook, M.J., & Neville, R.F. The faculty as teachers: A perspective on evaluation. Baltimore, MD: University of Maryland, Division of Education, 1971. (ERIC Document Reproduction Service No. ED 054-392)

Costin, R., Greenough, W.T., and Menges, R.J. Student ratings of college teaching: Reliability, validity, and usefulness. Review of Educational Research, 1971, 41, 511-533.

Crawford, P.L., & Bradshaw, H.L. Perception of characteristics of effective university teachers: A scale analysis. Educational and Psychological Measurement, 1968, 28, 1079-7085.

Cronbach, L.J. Essentials of psychological testing. (3rd. ed.). New York: Harper & Row, 1970.

Cronbach, L.J., & Meehl, P.E. Construct validity in psychological tests. Psychological Bulletin, 1955, 52, 281-302.

Dennis, L.J. Teacher evaluation in higher education. Liberal Education, 1976, 62, 437-443.

DeYoung, A.J. Classroom Climate and class success. The Journal of Educational Research, 1977, 252-257.

Doyle, K.O., Jr. Student evaluation of instruction. Lexington, MA: D.C. Heath, 1975.

Dressel, P.L. Handbook of academic evaluation. San Francisco: Jossey-Bass, 1976.

Elliott, D.N. Characteristics and relationships of various criteria of teaching. Thesis, Purdue University, 1949.

Equal Employment Opportunity Commission. Guidelines on employment selection procedures. Title 29, Chapter XIV, Part 1607 of the Code of Federal Regulations. Appearing in the Federal Register. 1970, 35 (149), 12333-12336.

Faia, M.A. How - and why - to cheat on student course evaluations. Liberal Education, 1976, 62, 113-119.

Ghiselli, E.E. Theory of psychological measurement. New York: McGraw-Hill, 1964.

Gordon, W.E., & Hudson W.W. Student evaluation of teachers and courses: What does it mean in social work. Paper presented at Annual Program Meeting - Council on Social Work Education, Phoenix, 1977. Prepared by Professors Gordon and Hudson of the George Warren Brown School of Social Work.

Grinnell, R.M. Jr., & Kyte, N.S. Measuring faculty competence. Journal of Education for Social Work, 1976, 12(3), 44-50.

Guion, R.M. Personnel Testing. New York: McGraw-Hill, 1965.

Gupta, Venue G. The effect of different instructions on student ratings of faculty. Washington: D.C. Paper presented to the annual convention of the American Psychological Association, 1976. (ERIC Document Reproduction Service No. 134 090)

Harari, O., & Zedeck, S. Development of behaviorally anchored scales for the evaluation of faculty teaching. Journal of Applied Psychology, 1973, 58, 261-265.

Holleman, I.T. Jr. The theory of evaluation with an emphasis on the formative phase. Document origin unknown: 1975. (ERIC Document Reproduction Service No. ED 129 855)

Isaacson, R.L., McKeachie, W.J., & Milholland, J.E. Correlation of teacher personality variables and student ratings. Journal of Educational Psychology, 1963, 54, 110-117.

Kerlinger, F.N. Foundations of behavioral research, (2nd ed.). New York: Holt, Rinehart and Winston, 1973.

Kirchner, R.P. A control factor in teacher evaluation by students. Unpublished research paper. Lexington, Kentucky: College Education, University of Kentucky, 1969.

Leventhal, L., Abrami, P.C., & Perry R.P. Do teacher rating forms reveal as much about students as about teachers? Journal of Educational Psychology, 1976, 68, 441-445.

Likert, R.A. A technique for the measurement of attitudes, Archives of Psychology, 1932, No. 140.

McKeachie, W.J. Research on teaching at the college and university level. In N.L. Gage (Eds.) Handbook on Teaching. Chicago: Rand McNally, 1963.

McKeachie, W.J. Student ratings of faculty. AAUP Bulletin, 1969, 55, 439-444.

McKeachie, W.J., Lin, Y., & Mann, W. Student ratings of teacher effectiveness. American Educational Research Journal, 1971, 8, 435-445.

Medley, D.M. Closing the gap between research teacher effectiveness and the teacher education curriculum. Journal of Research and and Development in Education, 7 (1) 1973, 12.

Meeth, R.L. The stateless art of teaching evaluation. Change Magazine. July 1976, 3-5.

Mehrens, W.A. & Ebel, R.L. (Eds.). Principles of educational and psychological measurement. Chicago: Rand McNally, 1967.

Millard, J.E. Teacher evaluation. Ankeny, IA: Heartland Education Agency, 1976. (ERIC Document Reproduction Service No. ED 132 195)

Miller, R.I. Evaluating faculty performance. San Francisco: Jossey-Bass, 1972.

Miller, R.I. Developing programs for faculty evaluation. San Francisco: Jossey-Bass. 1974.

Osgood, C.E., Suci, G.J., & Tannenbaum, P.H. The Measurement of meaning. Urbana, IL: University of Illinois Press, 1957.

Perry R.R. Evaluation of teaching behavior seeks to measure effectiveness, College and University Business, 1969, 47, 18-22.

Pyper, J.S. Critical elements in analysis and revision of instructional systems. Paper presented at California Educational Research Association Annual Meeting, 1971. (ERIC Document Reproduction Service No. ED 058 287)

Sheehan, D.S. On the invalidity of student ratings for administrative personnel decisions. The Journal of Higher Education, 1975, XLVI, 687-700.

Sherman, T.M. Trick or trait: A look at student evaluation of instruction. Educational Technology, 1976, 16 (4), 38-40.

Sorey, K.E. A study of the distinguishing personality charateristics of college faculty who are superior in regard to the teaching function. Unpublished doctoral dissertation, Oklahoma State University.

Stake, R.E. Overview and critique of existing evaluation practices and some new leads for the future. Urbana - Champaign, IL: University of Illinois, Center for Instructional Research and Curriculum Evaluation, 1976. (ERIC Document Reproduction Service No. ED 133 353)

Starkman, S.S., & Bellis, D.D. III. A novel approach to assessment of teacher preparation program effectiveness. Evaluation Quarterly, 1977, 1, 329-339.

Stone, E.F., Rabinowitz, S., & Spool, M.D. Effect of anonymity on student evaluations of faculty performance. Journal of Educational Psychology, 1977, 69, 274-280.

Sullivan, A.M. & Skanes, G.R. Validity of Student evaluation of teaching and the characteristics of successful instructors. Journal of Educational Psychology. 1974, 66, 584-590.

Swanson, R.A., & Sisson, D.J. The development, evaluation, and utilization of a departmental faculty appraisal system. Journal of Industrial Teacher Education, 1971, 9, 64-79.

Thomas, I.D. Considerations in a course evaluation in higher education, Educational Technology, 1976, 16 (4), 32-36.

Wilson, R.C., Gaff, J.G., Dienst, E.R., Wood, L., & Bavry, J.L. College professors and their impact on students. New York: Wiley, 1975.

Winer, B.J. Statistical principles in experimental design. (2nd ed.). New York: McGraw-Hill, 1971.

Yongkittikul, C., Gillmore, G., & Brandenburg, D.C. Does the time of course meeting affect course ratings by students? (RR 346). Urbana, IL: University of Illinois, Measurement and Research Division, Office of Instructional Resources, 1974.